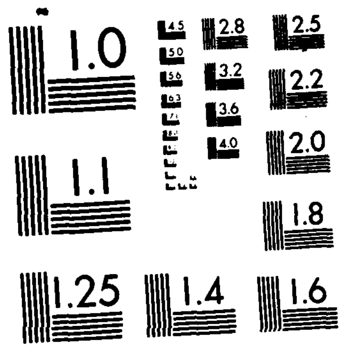
AD-A164 638 MEASURES OF IMBALANCE FOR UNBALANCED MODELS(U) FLORIDA UNIV GAINEVVILLE DEPT OF STATISTICS A I KHURI JAN 86 TR-255 N00014-86-K-0059 1/1

UNCLASSIFIED F/G 12/1 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(12)

MEASURES OF IMBALANCE FOR UNBALANCED MODELS

By

Andre' I. Khuri

Department of Statistics, University of Florida
Gainesville, FL. 32611

Technical Report Number 255
January, 1986

PREPARED UNDER GRANT NO. N00014-86-K-0059 FROM THE OFFICE OF NAVAL RESEARCH. ANDRE' I. KHURI AND RAMON C. LITTELL, PRINCIPAL INVESTIGATORS.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 255	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): MEASURES OF IMBALANCE FOR UNBALANCED MODELS	5. TYPE OF REPORT & PERIOD COVERED
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Andre' I. Khuri	8. CONTRACT OR GRANT NUMBER(s): N00014-86-K-0059 (R&T 4114552---01)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, Nuclear Sciences Center, University of Florida, Gainesville, FL 32611	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Mathematical Sciences Division (Code 411), Arlington, VA 22217-5000	12. REPORT DATE: January, 1986
	13. NUMBER OF PAGES: 21
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Unbalanced data, Nested models, Cross-classification models, Loglinear models, Chi-squared statistic.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Loglinear models are used to measure the degree of imbalance of an unbalanced data set. The proposed procedure can also be used to measure departures from certain types of balance, such as proportionality of subclass frequencies, partial balance, and last-stage uniformity.

Measures of Imbalance For

Unbalanced Models

A. I. Khuri

The University of Florida

Technical Report Number 255

Department of Statistics
University of Florida
Gainesville, Florida 32611

January 1986

Measures of Imbalance For
Unbalanced Models

A.I. Khuri

The University of Florida

Abstract

In this paper we present a procedure to measure the degree of imbalance of an unbalanced data set. The procedure is based on choosing an appropriate loglinear model for the subclass frequencies of the data. A measure of imbalance is then introduced as some function of the chi-squared statistic used in the goodness-of-fit test for the loglinear model. The proposed procedure can also be used to measure departures from certain types of balance, such as proportionality of subclass frequencies, partial balance, and last-stage uniformity.

Key words: Unbalanced data, nested models, cross-classification models, loglinear models, chi-squared statistic.

1. Introduction

It is known that in a balanced data situation, parameter estimators and test statistics pertaining to the effects in the associated model have certain optimal properties. These properties, however, cannot be maintained once the data set becomes unbalanced. In this case, the statistical properties of the aforementioned estimators and test statistics will, to a large extent, depend on the pattern of the data subclass frequencies. Severe imbalance in the data can have adverse effects on the analysis, especially if that analysis is an adaptation of procedures pertaining to balanced data (see, for example,

Cummings and Gaylor 1974).

Ahrens and Pincus (1981) presented two measures of imbalance for the one-way classification model. These measures were utilized to assess the efficiency of an associated unbalanced design as compared to a balanced design with the same number of observations. Other authors have alluded to the need to measure data imbalance; they include Hess (1979, p. 646) and Tietjen (1974, p. 576).

The purpose of this paper is to present a general procedure to measure imbalance of a data set for a given unbalanced model. It is shown that one of the two measures introduced by Ahrens and Pincus (1981) can be derived as a special case using this procedure. The proposed procedure can also be utilized to measure departures from certain types of balance other than complete balance where frequencies are equal within all the subclasses. These include partial balance, last-stage uniformity, and the case of proportional subclass frequencies in cross-classification models.

2. A General Procedure to Measure Imbalance

A measure of imbalance, denoted by $\phi(D)$, is a function of the subclass frequencies which are determined by the design D used in the experiment. This function takes values inside the closed interval $[0,1]$. Small values of $\phi(D)$ indicate severe imbalance, whereas "near balance" cases are characterized by large values of $\phi(D)$. The data set is balanced if and only if $\phi(D)=1$. Furthermore, this function must remain invariant under any partial or complete replication of the design (see Ahrens and Pincus 1981).

The development of the function $\phi(D)$ is based on the use of loglinear models. Several unbalanced models will be considered to illustrate the application of this procedure. These models include the one-way classifaction model, the two-way classification model, the three-way classification model,

the two-fold nested model, the three-fold nested model, and a model with a mixture of cross-classified and nested effects.

2.1 The One-Way Classification Model

Consider the one-way model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{2.1}$$

($i = 1,2,\ldots,a$; $j = 1,2,\ldots,n_i$), where μ is a fixed unknown parameter, α_i is either a fixed parameter or a random variable, and ε_{ij} is a random error. Here D is the design $D = \{n_1,n_2,\ldots,n_a\}$.

We shall consider that the n_i's have a multinomial distribution such that n_i has the binomial distribution $B(n_{\cdot},\Pi_i)$, where $n_i = \Sigma_{i=1}^{a} n_i$ and Π_i is the probability of belonging to level i ($i = 1,2,\ldots,a$). Hence, $m_i = E(n_i) = n_{\cdot}\Pi_i$ ($i = 1,2,\ldots,a$). The m_i's will be referred to as expected frequencies. On a logarithmic scale, the expected frequencies can be represented by the loglinear model

$$\log m_i = \bar{\mu} + \bar{\alpha}_i, \quad i = 1,2,\ldots,a, \tag{2.2}$$

where

$$\bar{\mu} = \log n_{\cdot} + \frac{1}{a}\sum_{i=1}^{a} \log \Pi_i$$

$$\bar{\alpha}_i = \log \Pi_i - \frac{1}{a}\sum_{i=1}^{a} \log \Pi_i, \quad i = 1,2,\ldots,a.$$

We note that $\Sigma_{i=1}^{a}\bar{\alpha}_i = 0$ and that model (2.2) is of the same form as model (2.1), except for the error term.

Let $\hat{m}_i$ denote the maximum likelihood estimate of m_i ($i = 1,2,\ldots,a$). Under complete balance, $\Pi_i = 1/a$ for all i, hence $\hat{m}_i = n_{\cdot}/a = \bar{n}_{\cdot}$. Using Pearson's approximate chi-squared statistic for testing the hypothesis $H_o:\Pi_i = 1/a$ for all i we obtain

$$X^2 = \sum_{i=1}^{a} (n_i-\bar{n}_.)^2/\bar{n}_.,$$

which under H_o has an asymptotic chi-squared distribution with θ degrees of freedom, where, in general, θ is the difference between the number of independent Π_i's under H_a and under H_o, respectively. In this case $\theta = a-1$. We define our measure of imbalance as

$$\phi(D) = \frac{1}{1+c^2}, \tag{2.3}$$

where $c^2 = X^2/n_.$. We note that $0 \leq \phi(D) \leq 1$ and the division of X^2 by $n_.$ causes the measure to be invariant to any replication of the design as required. Furthermore, $\phi(D) = 1$ if and only if the n_i's are equal. We also note that $\phi(D)$ is identical to the measure $\nu(D)$ given by Ahrens and Pincus (1981).

2.2 The Two-Way Classification Model

Consider the model

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \tag{2.4}$$

$(i = 1,2,\ldots,a;\ j = 1,2,\ldots,b;\ k = 1,2,\ldots,n_{ij})$, where μ is a fixed unknown parameter; α_i and β_j can be either fixed or random. In this case the design D is $D = \{n_{11},n_{12},\ldots,n_{ab}\}$. The n_{ij}'s are considered to have the multinomial distribution and each n_{ij} has the binomial distribution $B(n_{..},\Pi_{ij})$, where $n_{..} = \Sigma_{i,j}n_{ij}$ and Π_{ij} is the probability of belonging to the $(i,j)^{th}$ cell. Hence, $E(n_{ij}) = m_{ij} = n_{..}\Pi_{ij}$. The corresponding loglinear model is

$$\log m_{ij} = \bar{u} + \bar{\alpha}_i + \bar{\beta}_j + (\overline{\alpha\beta})_{ij}, \tag{2.5}$$

where in this case

$$\begin{aligned}
\bar{u} &= \frac{1}{ab}\sum_{i,j} \log m_{ij} \\
\bar{\alpha}_i &= \frac{1}{b}\sum_{j} \log m_{ij} - \bar{\mu} \\
\bar{\beta}_j &= \frac{1}{a}\sum_{i} \log m_{ij} - \bar{\mu}
\end{aligned} \tag{2.6}$$

$$(\overline{\alpha\beta})_{ij} = \log m_{ij} - \frac{1}{a}\sum_i \log m_{ij} - \frac{1}{b}\sum_j \log m_{ij} + \bar{\mu} .$$

We note that models (2.4) and (2.5) are of the same form, except for the error term. The $\bar{\alpha}_i$'s, $\bar{\beta}_j$'s and $(\overline{\alpha\beta})_{ij}$'s satisfy

$$\sum_i \bar{\alpha}_i = \sum_j \bar{\beta}_j = \sum_i (\overline{\alpha\beta})_{ij} = \sum_j (\overline{\alpha\beta})_{ij} = 0.$$

Let $\hat{m}_{ij}$ denote the maximum likelihood estimate of m_{ij} $(i = 1,2,\ldots,a;\ j = 1,2,\ldots,b)$. Under the hypothesis $H_o{:}\Pi_{ij} = \Pi_i\Pi_j$ for all i and j (this is called the hypothesis of independence), where $\Pi_i = \Sigma_j\Pi_{ij}$ and $\Pi_j = \Sigma_i\Pi_{ij}$, the maximum likelihood estimates of Π_i and Π_j are $n_{i\cdot}/n_{\cdot\cdot}$ and $n_{\cdot j}/n_{\cdot\cdot}$, respectively, where $n_{i\cdot} = \Sigma_j n_{ij}$ and $n_{\cdot j} = \Sigma_i n_{ij}$. Hence, $\hat{m}_{ij} = n_{i\cdot}\, n_{\cdot j}/n_{\cdot\cdot}$. This is the case of proportional subclass frequencies. The corresponding test statistic is

$$X^2 = \sum_{i,j} (n_{ij} - \hat{m}_{ij})^2/\hat{m}_{ij},$$

which under H_o has an asymptotic chi-squared distribution with $\theta = (a-1)(b-1)$ degrees of freedom. If $c^2 = X^2/n_{\cdot\cdot}$, then

$$\phi(D) = \frac{1}{1+c^2} \qquad (2.7)$$

is a measure of departure from proportionality of the subclass frequencies with $\phi(D)$ attaining the value one when these frequencies are proportional. In the latter case, model (2.5) takes the additive form

$$\log m_{ij} = \bar{\mu} + \bar{\alpha}_i + \bar{\beta}_j . \qquad (2.8)$$

Under the hypothesis of complete balance, namely, $H_o{:}\Pi_{ij} = 1/(ab)$ for all i and j, $\hat{m}_{ij} = n_{\cdot\cdot}/(ab)$, and the corresponding statistic,

$$X^2 = \sum_{i,j} \frac{\left[n_{ij} - n_{\cdot\cdot}/(ab)\right]^2}{n_{\cdot\cdot}/(ab)}, \qquad (2.9)$$

is asymptotically distributed as a chi-squared variate with $\theta = ab - 1$ degrees of freedom. A measure of departure from complete balance is then given by

$$\phi(D) = \frac{1}{1+c^2} , \qquad (2.10)$$

where $c^2 = X^2/n_{..}$. In this case model (2.8) is reduced to just

$$\log m_{ij} = \bar{\mu} .$$

2.3 The Three-Way Classification Model

Suppose we consider the model

$$y_{ijk\ell} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijk\ell} \qquad (2.11)$$

($i = 1,2,\ldots,a$; $j = 1,2,\ldots,b$; $k = 1,2,\ldots,c$; $\ell = 1,2,\ldots,n_{ijk}$), α_i, β_j, and γ_k can be either fixed or random. The design D consists of the cell frequencies, $n_{111}, n_{112}, \ldots, n_{abc}$. Following the approach used in the earlier two models, if $m_{ijk} = E(n_{ijk})$, then $\log m_{ijk}$ can be expressed in terms of the loglinear model

$$\log m_{ijk} = \bar{\mu} + \bar{\alpha}_i + \bar{\beta}_j + \bar{\gamma}_k + (\overline{\alpha\beta})_{ij} + (\overline{\alpha\gamma})_{ik} + (\overline{\beta\gamma})_{jk} + (\overline{\alpha\beta\gamma})_{ijk} , \qquad (2.12)$$

where

$$\sum_i \bar{\alpha}_i = \sum_j \bar{\beta}_j = \sum_k \bar{\gamma}_k = \sum_i (\overline{\alpha\beta})_{ij} = \sum_j (\overline{\alpha\beta})_{ij} = \ldots = \sum_k (\overline{\alpha\beta\gamma})_{ijk} = 0 .$$

From (2.12) several reduced models may be considered. These models are given in Table 1. The goodness-of-fit of these models can be checked by using Pearson's approximate chi-squared statistic

$$X^2 = \sum_{i,j,k} (n_{ijk} - \hat{m}_{ijk})^2 / \hat{m}_{ijk} , \qquad (2.13)$$

where $\hat{m}_{ijk}$ is the maximum likelihood estimate of m_{ijk}, or by the likelihood ratio statistic

$$G^2 = 2 \sum_{i,j,k} n_{ijk} \log(n_{ijk}/\hat{m}_{ijk}) \qquad (2.14)$$

(see Agresti 1984, p. 48). Both X^2 and G^2 are asymptotically distributed as chi-squared variates with the same degrees of freedom. The $\hat{m}_{ijk}$ estimates for the models in Table 1 are given in the same table along with the corresponding

Table 1

Some Loglinear Models For a Three-Factor Experiment

	Model	$\hat{m}_{ijk}$**	X^2	G^2	Degrees of Freedom
I.	$\log m_{ijk} = \bar{\mu} + \bar{\alpha}_i + \bar{\beta}_j + \bar{\gamma}_k + (\overline{\alpha\beta})_{ij} + (\overline{\alpha\gamma})_{ik} + (\overline{\beta\gamma})_{jk} + (\overline{\alpha\beta\gamma})_{ijk}$	n_{ijk}	$X_1^2 = 0$	$G_1^2 = 0$	0
II.*	$\log m_{ijk} = \bar{\mu} + \bar{\alpha}_i + \bar{\beta}_j + \bar{\gamma}_k + (\overline{\alpha\beta})_{ij} + (\overline{\beta\gamma})_{jk}$	$\frac{n_{ij\cdot}n_{\cdot jk}}{n_{\cdot j\cdot}}$	X_2^2	G_2^2	b(a-1)(c-1)
III.*	$\log m_{ijk} = \bar{\mu} + \bar{\alpha}_i + \bar{\beta}_j + \bar{\gamma}_k + (\overline{\alpha\beta})_{ij}$	$\frac{n_{ij\cdot}n_{\cdot\cdot k}}{n_{\cdot\cdot\cdot}}$	X_3^2	G_3^2	(ab-1)(c-1)
IV.	$\log m_{ijk} = \bar{\mu} + \bar{\alpha}_i + \bar{\beta}_j + \bar{\gamma}_k$	$\frac{n_{i\cdot\cdot}n_{\cdot j\cdot}n_{\cdot\cdot k}}{n_{\cdot\cdot\cdot}^2}$	X_4^2	G_4^2	abc-a-b-c+2
V.	$\log m_{ijk} = \bar{\mu}$	$\frac{n_{\cdot\cdot\cdot}}{abc}$	X_5^2	G_5^2	abc-1

* Two other models of similar form exist.
** The dots used in the elements of this column denote summation over the corresponding missing subscripts.

X^2 and G^2 statistics and associated degrees of freedom. The G^2 statistic has the desirable feature of being monotone increasing as terms are deleted from the full model in (2.12), that is, $0 = G_1^2 \leq G_2^2 \leq G_3^2 \leq G_4^2 \leq G_5^2$ (see Agresti 1984, p. 57). The G^2 statistic can, therefore, be used to compare two nested models (that is, one model is obtained from the other by deleting one or more terms) that give adequate fits to the cell frequencies. Thus, with the help of the G^2 statistic it is possible to identify one or more models in Table 1 that provide adequate fits. For such models departures of cell frequencies from their expected values can be measured by means of the function $\phi(D)$ in (2.10) where c^2 is given by the corresponding value of X^2 in Table 1 divided by $n_{...}$.

Model V in Table 1 corresponds to the case of complete balance, whereas Model IV is associated with the case of proportional subclass frequencies. Model III corresponds to a case of conditional proportional subclass frequencies involving values of i, k for a fixed j, and values of j and k for a fixed i. In Model II we have a case of conditional proportional subclass frequencies involving only values of i and k for a fixed j. Model I is the full loglinear model.

2.4 The Two-Fold Nested Classification Model

Let us now consider the model

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk} \tag{2.15}$$

($i = 1,2,\ldots,a$; $j = 1,2,\ldots,b_i$; $k = 1,2,\ldots,n_{ij}$), α_i denotes the nesting effect and β_{ij} denotes the nested effect. The design D consists of the values of b_1, b_2, ..., b_a in addition to the n_{ij} values. In the complete balance case $b_i=b$ for all i and $n_{ij}=n$ for all i and j. A condition weaker than complete balance is last-stage uniformity which requires that $n_{ij}=n$ for all i and j. If, however, $n_{ij}=n_{ij'}$ for $j \neq j'$ and $i = 1,2,\ldots,a$, then the design is partially balanced. It is known that when all the effects in (2.15) are random, last-stage

uniformity is a sufficient condition for the sums of squares, in the conventional analysis of variance table, to be independently distributed as scaled chi-squared variates (see Tietjen 1974, p. 575). Under partial balance, however, the sums of squares for the α_i and β_{ij} effects are independent, but do not have the scaled chi-squared distribution (see Cummings 1972). It is, therefore, of interest to measure departures from complete balance, last-stage uniformity, and partial balance.

The loglinear model corresponding to model (2.15) can be obtained as follows: let m_{ij} denote the expected frequency $E(n_{ij})$. Then, $m_{ij} = n_{..}\Pi_{ij} = n_{..}\Pi_i\Pi_{j|i}$, where Π_i denotes the probability of belonging to the i^{th} level of the nesting factor, $\Pi_{j|i}$ denotes the conditional probability of belonging to a j^{th} level of the nested factor given the i^{th} level of the nesting factor. Hence, $\log m_{ij}$ can be represented by the loglinear model

$$\log m_{ij} = \bar{\mu} + \bar{\alpha}_i + \bar{\beta}_{ij} \ , \tag{2.16}$$

where

$$\bar{\mu} = \log n_{..} + \frac{1}{b_.}\sum_{i=1}^{a} b_i \log \Pi_i + \frac{1}{b_.}\sum_{i=1}^{a}\sum_{j=1}^{b_i} \log \Pi_{j|i}$$

$$\bar{\alpha}_i = \log \Pi_i + \frac{1}{b_i}\sum_{j=1}^{b_i} \log \Pi_{j|i} - \frac{1}{b_.}\sum_{i=1}^{a} b_i \log \Pi_i - \frac{1}{b_.}\sum_{i=1}^{a}\sum_{j=1}^{b_i} \log \Pi_{j|i}$$

$$\bar{\beta}_{ij} = \log \Pi_{j|i} - \frac{1}{b_i}\sum_{j=1}^{b_i} \log \Pi_{j|i} \ .$$

In the partial balance case, $\Pi_{j|i} = 1/b_i$ for all i and j, hence, the maximum likelihood estimate of m_{ij} is $\hat{m}_{ij} = n_{i.}/b_i$, where $n_{i.} = \Sigma_{j=1}^{b_i} n_{ij}$ $(i = 1,2,\ldots,a)$, since in this case $\hat{\Pi}_i = n_{i.}/n_{..}$. A measure of partial balance is then given by

$$\phi(D) = \frac{1}{1+c^2} \tag{2.17}$$

where

$$c^2 = X^2/n_{\cdot\cdot}$$

and

$$X^2 = \sum_{i,j} \frac{(n_{ij} - n_{i\cdot}/b_i)^2}{n_{i\cdot}/b_i} \quad . \tag{2.18}$$

Under partial balance X^2 has asymptotically the chi-squared distribution with $b_\cdot - a$ *degrees of freedom*, where $b_\cdot = \Sigma_{i=1}^{a} b_i$. This follows from the fact that in the general case, the number of linearly independent Π_{ij}'s is $b_\cdot - 1$ whereas under partial balance this number is just a-1. We note that this case can be represented by the loglinear model

$$\log m_{ij} = \bar{\mu} + \bar{\alpha}_i \; . \tag{2.19}$$

Under last-stage uniformity, $\Pi_{ij} = 1/b_\cdot$ for all i and j. The loglinear model in this case has the form

$$\log m_{ij} = \bar{\mu} \quad . \tag{2.20}$$

The maximum likelihood estimate of m_{ij} is given by $\bar{n}_{\cdot\cdot} = n_{\cdot\cdot}/b_\cdot$. Hence, a measure of departure from last-stage uniformity is given by (2.17), where

$$c^2 = X^2/n_{\cdot\cdot}$$

and

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \bar{n}_{\cdot\cdot})^2}{\bar{n}_{\cdot\cdot}} \; , \tag{2.21}$$

which has the asymptotic chi-squared distribution with $b_\cdot - 1$ degrees of freedom.

Unlike the former two cases, departure from complete balance can be attributed to variation in the values of $b_1, b_2, \ldots, b_a$, or to variation in the n_{ij} values. We thus need to measure imbalance with regard to the b_i's and also with regard to the n_{ij}'s. We shall consider that the b_i's form a multinomial distribution independently of the multinomial distribution of the n_{ij}'s, with b_i

being distributed as a binomial $B(b_{\cdot},\tau_i)$. Hence, $d_i = E(b_i) = b_{\cdot}\tau_i$ $(i = 1,2,\ldots,a)$. A measure of imbalance concerning the b_i's is, therefore, given by

$$\phi_1(D) = \frac{1}{1+c_1^2} , \tag{2.22}$$

where

$$c_1^2 = X_1^2/b_{\cdot}$$

and

$$X_1^2 = \sum_{i=1}^{a} \frac{(b_i-\bar{b}_{\cdot})^2}{\bar{b}_{\cdot}} , \tag{2.23}$$

where $\bar{b}_{\cdot} = b_{\cdot}/a$. This statistic has the asymptotic chi-squared distribution with a-1 degrees of freedom when $\tau_i = 1/a$ $(i = 1,2,\ldots,a)$. On the other hand, a measure of imbalance concerning the n_{ij}'s is

$$\phi_2(D) = \frac{1}{1+c_2^2} , \tag{2.24}$$

where

$$c_2^2 = X_2^2/n_{\cdot\cdot}$$

and

$$X_2^2 = \sum_{i,j} \frac{(n_{ij}-\bar{n}_{\cdot\cdot})^2}{\bar{n}_{\cdot\cdot}} . \tag{2.25}$$

The statistic X_2^2 is the same as the one used in last-stage uniformity. Since the multinomial distribution of the b_i's is independent of the multinomial distribution of the n_{ij}'s, X_1^2 is statistically independent of X_2^2, hence $X_1^2 + X_2^2$ is asymptotically distributed as a chi-squared variate with $b_{\cdot}$+a-2 degrees of freedom. Now, to measure departure from complete balance we use the measure

$$\phi(D) = \frac{1}{1+c^2} , \tag{2.26}$$

where

$$c^2 = c_1^2 + c_2^2 \tag{2.27}$$

2.5 The Three-Fold Nested Classification Model

In this section we consider the model

$$y_{ijk\ell} = \mu + \alpha_i + \beta_{ij} + \gamma_{ijk} + \varepsilon_{ijk\ell} \quad , \tag{2.28}$$

$(i = 1,2,\ldots,a;\ j = 1,2,\ldots,b_i;\ k = 1,2,\ldots,c_{ij};\ \ell = 1,2,\ldots,n_{ijk})$. The values of b_i, c_{ij}, and n_{ijk} make up the design D. Here different types of balance can be considered; each is a stronger type of balance than the one preceding it:

i) Last-stage partial balance, that is, partial balance with respect to the n_{ijk} values. There are two kinds of such partial balance; in the first kind n_{ijk} depends on i and j only, and in the second kind n_{ijk} depends on i only.

ii) Last-stage uniformity when the n_{ijk}'s are equal for all values of i,j, and k.

iii) Last-stage uniformity and next-to-last-stage partial balance, that is, when c_{ij} depends on i only.

iv) Last-stage uniformity as well as next-to-last-stage uniformity, that is, when the c_{ij}'s are equal for all values of i and j.

v) Complete balance. This occurs when equality of frequencies occurs within all the subclasses.

Each type can be characterized by one or more loglinear models and a corresponding measure of imbalance can be obtained accordingly. For example, for Type (iii), if the n_{ijk}'s are considered to have a multinomial distribution with n_{ijk} being distributed as a binomial $B(n_{...},\pi_{ijk})$, where $n_{...} = \sum_{i,j,k} n_{ijk}$,

then $m_{ijk} = E(n_{ijk}) = n_{...}/c_{..}$ under last-stage uniformity. Furthermore, if the c_{ij}'s have a multinomial distribution, independently of the n_{ijk}'s, with c_{ij} distributed as $B(c_{..},\tau_{ij})$, where $c_{..} = \sum_{i,j} c_{ij}$, then $d_{ij} = E(c_{ij}) = c_{..}\tau_i/b_i$ $(i = 1,2,\ldots,a)$, since under partial balance with respect to the c_{ij}'s, $\tau_{ij} = \tau_i\tau_{j|i} = \tau_i/b_i$. Thus, the associated loglinear models for m_{ijk} and d_{ij} are

$$\log m_{ijk} = \bar{\mu}_1 \quad ,$$

$$\log d_{ij} = \bar{\mu}_2 + \bar{\alpha}_i \quad .$$

A measure of imbalance for Type (iii) balance is, therefore, given by

$$\phi(D) = \frac{1}{1+c^2} \quad ,$$

where

$$c^2 = c_1^2 + c_2^2$$

and where

$$c_1^2 = \frac{1}{n_{...}} \sum_{i,j,k} \frac{(n_{ijk}-n_{...}/c_{..})^2}{n_{...}/c_{..}} \quad , \qquad (2.29)$$

$$c_2^2 = \frac{1}{c_{..}} \sum_{i,j} \frac{(c_{ij}-c_{i.}/b_i)^2}{c_{i.}/b_i} \quad , \qquad (2.30)$$

since in this case the maximum likelihood estimate of d_{ij} is $\hat{d}_{ij} = c_{..}\hat{\tau}_i/b_i$ $= c_{..}(c_{i.}/c_{..})/b_i = c_{i.}/b_i$. We note that $n_{...}c_1^2$ and $c_{..}c_2^2$ are distributed independently as asymptotic chi-squared variates with $c_{..}-1$ and $b_.-1$ degrees of freedom, respectively.

2.6 A Model With A Mixture Of Cross-Classified And Nested Effects

Consider a model involving three factors, A, B, and C, with A and C crossed and B is nested within A. This model is written as

$$y_{ijk\ell} = \mu + \alpha_i + \beta_{ij} + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{ijk} + \varepsilon_{ijk\ell} \qquad (2.31)$$

($i = 1,2,\ldots,a$; $j = 1,2,\ldots,b_i$; $k = 1,2,\ldots,c$; $\ell = 1,2,\ldots,n_{ijk}$). Let Π_{ijk} denote the probability of belonging to level i of A, level j of B nested within i, and level k of C. As before, the n_{ijk}'s are considered to have a multinomial distribution with $m_{ijk} = E(n_{ijk}) = n_{\cdots}\Pi_{ijk}$.

For this model we can have four types of balance:

i) Proportional subclass frequencies involving the AB subclasses and the levels of factor C, that is, $\Pi_{ijk} = \Pi_i \Pi_{j|i} \Pi_k$.

ii) Partial balance with respect to the n_{ijk} values, that is, $\Pi_{ijk} = \Pi_i/(b_i c)$.

iii) Last-stage uniformity, that is, $\Pi_{ijk} = 1/(b_. c)$ for all i, j, and k.

iv) Complete balance, that is, $\tau_i = 1/a$ and $\Pi_{ijk} = 1/(b_. c)$, where τ_i is the i^{th} binomial probability associated with the multinomial distribution of the b_i's.

Each of the above four types can be represented by a *loglinear model*. These models are given in Table 2. Furthermore, for each of these four types a measure of imbalance is obtained by using formula (2.3). The value of c^2 in this formula and the degrees of freedom for the corresponding asymptotic chi-squared statistics are also given in Table 2.

3. Numerical Examples

i) Cummings and Gaylor (1974) used several designs to illustrate the combined effects of dependence and nonchi-squaredness of the analysis of variance mean squares on the size of Satterthwaite's approximate F-test for variance component testing in a two-fold nested model. We shall consider three of these designs which are described in Table 3 and are also represented graphically in Figure 1.

For each of the three designs we measure departures from partial balance

Table 2

Loglinear Models Associated With Four Types Of Balance For Model (2.31)

Type	Loglinear Model	c^2	Degrees of Freedom
(i)	$\log m_{ijk} = \bar{\mu} + \bar{\alpha}_i + \bar{\beta}_{ij} + \bar{\gamma}_k$	$c_1^2 = \frac{1}{n_{...}} \sum_{i,j,k} \frac{(n_{ijk} - n_{ij.} n_{..k}/n_{...})^2}{n_{ij.} n_{..k}/n_{...}}$	$(b_. - 1)(c-1)$
(ii)	$\log m_{ijk} = \bar{\mu} + \bar{\alpha}_i$	$c_2^2 = \frac{1}{n_{...}} \sum_{i,j,k} \frac{\left[n_{ijk} - n_{i..}/(b_i c)\right]^2}{n_{i..}/(b_i c)}$	$b_. c - a$
(iii)	$\log m_{ijk} = \bar{\mu}$	$c_3^2 = \frac{1}{n_{...}} \sum_{i,j,k} \frac{\left[n_{ijk} - n_{...}/(b_. c)\right]^2}{n_{...}/(b_. c)}$	$b_. c - 1$
(iv)	$\log m_{ijk} = \bar{\mu}$ and $\log d_i = \bar{\mu}'$, where $d_i = E(b_i)$	$c_4^2 = c_3^2 + \frac{1}{b_.} \sum_i \frac{(b_i - b_./a)^2}{b_./a}$	$b_. c + a - 2$

and last-stage uniformity by using formula (2.17) with X^2 being given by (2.18) for partial balance and by (2.25) for last-stage uniformity. We also measure departure from complete balance by applying formulas (2.26) and (2.27). The results are given in Table 4.

Table 3

Designs For a Two-Fold Nested Model

		i = 1	2	3	4
Design 1	b_i	1	1	4	4
	n_{ij}	1	4	1,1,1,1	4,4,4,4
Design 2	b_i	2	2	2	2
	n_{ij}	1,5	1,5	1,5	1,5
Design 3	b_i	1	2	2	4
	n_{ij}	1	1,4	1,8	1,2,3,4

Table 4

Values of $\phi(D)$ For The Three Designs In Table 3

Design	Partial Balance	Last-Stage Uniformity	Complete Balance
1	1	.735	.58
2	.69	.69	.69
3	.73	.61	.54

From Table 4 we note that other than partial balance for Design 1, none of the designs has strong balance properties. Of all three designs, Design 3 is the most unbalanced with respect to last-stage uniformity and complete balance.

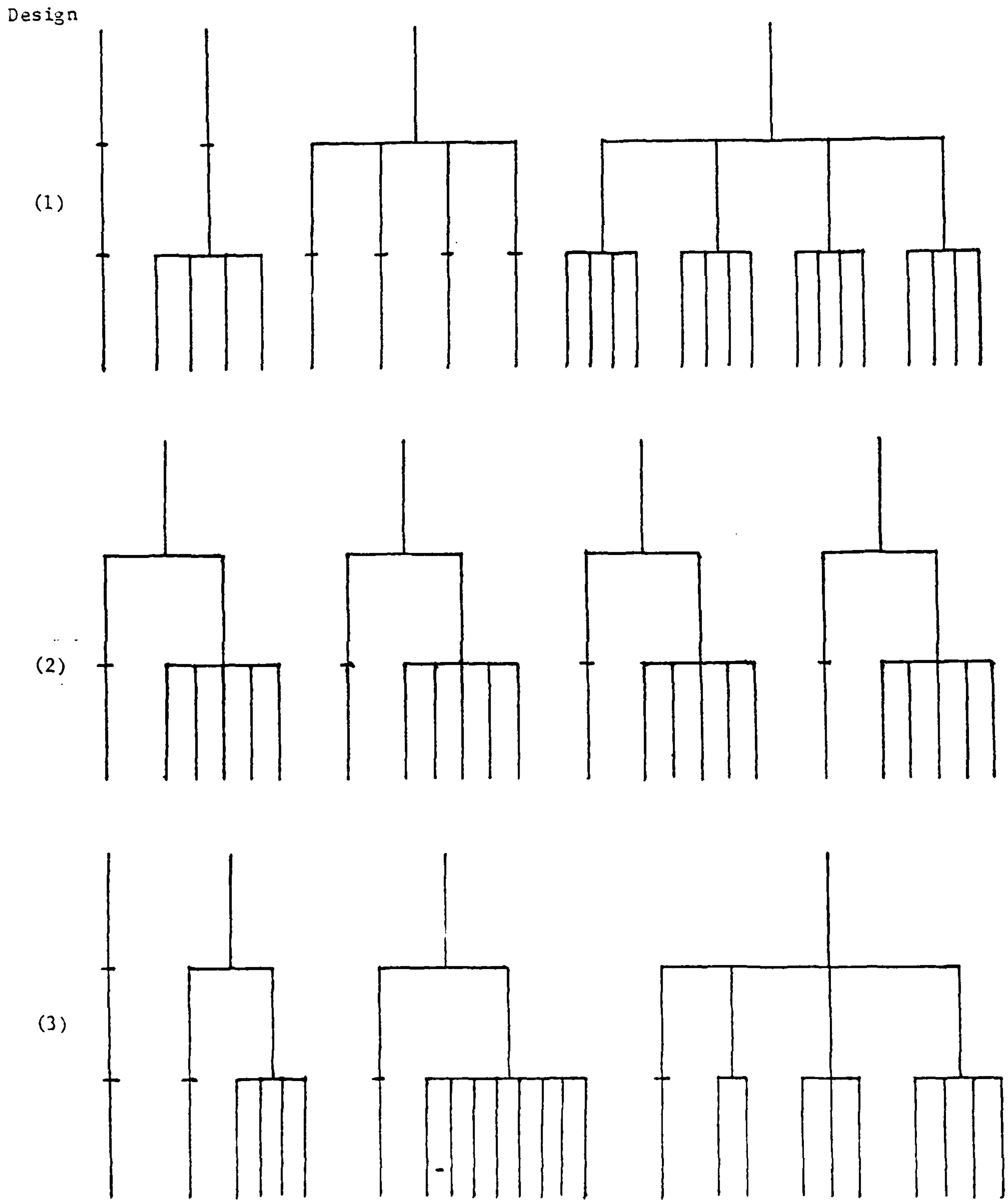


Fig. 1. Three Designs For a Two-Fold Nested Model

ii) Bliss (1967, p. 355) described a nested experiment involving three factors with an associated model of the form given by (2.28). In this experiment, a=11 and the design D consists of the following elements: $b_i=3$ ($i = 1,2,\ldots,11$); $c_{i1}=2$, $c_{i2}=2$, $c_{i3}=1$ ($i = 1,2,\ldots,11$); $n_{i11}=2$, $n_{i12}=2$, $n_{i21}=1$, $n_{i22}=1$, $n_{i31}=1$ ($i = 1,2,\ldots,11$). Graphically, for each value of i, the design D can be depicted as in Figure 2.

The design D is partially balanced of the first kind (see Section 2.5). A measure of departure from Type (ii) balance is given by

$$\phi(D) = \frac{1}{1+c_1^2},$$

where c_1^2 is described in (2.29), hence $\phi(D) = .89$. The measure for Type (iii) balance is given by

$$\phi(D) = \frac{1}{1+c_1^2+c_2^2},$$

where c_2^2 is described in (2.30), hence $\phi(D) = .83$. As for Type (iv) balance, the corresponding measure is

$$\phi(D) = \frac{1}{1+c_1^2+c_3^2},$$

where

$$c_3^2 = \frac{1}{c_{..}} \sum_{i,j} \frac{(c_{ij}-\bar{c}_{..})^2}{\bar{c}_{..}},$$

where $\bar{c}_{..} = c_{..}/b_.$, hence, $\phi(D) = .83$. We note that this is equal to the previous measure value for Type (iii) since both $c_{i.}$ and b_i in formula (2.30) do not depend on i, thus, $c_{i.}/b_i = c_{..}/b_. = \bar{c}_{..}$. We also note that since the b_i's are equal, the value $\phi(D) = .83$ is also a measure of departure from complete balance, which is Type (v) balance.

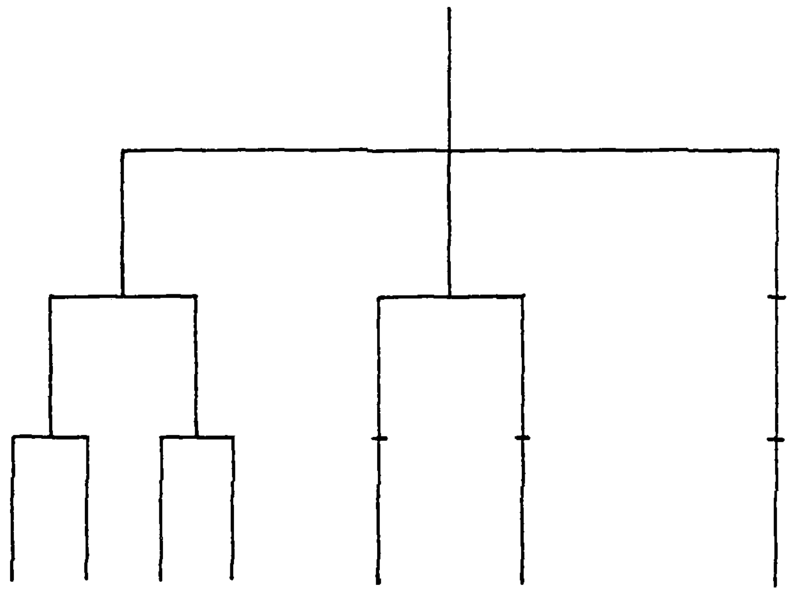


Fig. 2. A Design For a Three-Fold Nested Model

4. Concluding Remarks

We have introduced a procedure for measuring the degree of imbalance that is associated with an unbalanced model. The procedure applies to cross classification models, nested classification models, and to models with a mixture of cross-classified and nested effects. It can also be used to measure departures from different types of balance, especially in nested models where imbalance can affect various stages of the nested design. Several examples of unbalanced models were studied. From these examples it is easy to see that this procedure is general enough to apply to any unbalanced model.

With the help of this procedure it is now possible to describe in a quantitative manner different kinds of imbalance, such as extreme imbalance, moderate imbalance, and near balance. This can serve as an indicator of the suitability of the approximate methods that are adapted from balanced-data-based procedures and used to analyze an unbalanced model, particularly, when the appropriate measure value is near unity. It is to be cautioned, however, that low values of that measure do not necessarily mean that such approximate methods are inadequate. Cummings and Gaylor (1974), for example, noted that for some extremely unbalanced design, namely, Design 3 in Table 3, their approximate F-test performed very well. They attributed this behavior to counter-balancing effects which appear to reduce, rather than compound, the effect of imbalance on the standard analysis of variance.

References

Agresti, A., 1984: Analysis of Ordinal Categorical Data. Wiley, New York.

Ahrens, H. and R. Pincus, 1981: On two measures of unbalancedness in a one-way model and their relation to efficiency. Biom. J. 23, 227-235.

Bliss, C.I., 1967: Statistics in Biology, Volume 1. McGraw-Hill, New York.

Cummings, W.B., 1972: Variance component testing in unbalanced nested designs. North Carolina State University Institute of Statistics Mimeo Series No. 843.

Cummings, W.B. and D.W. Gaylor, 1974: Variance component testing in unbalanced nested designs. J. Amer. Statist. Assoc. 69, 765-771.

Hess, J.L., 1979: Sensitivity of MINQUE with respect to a priori weights. Biometrics 35, 645-649.

Tietjen, G.L., 1974: Exact and approximate tests for unbalanced random effects designs. Biometrics 30, 573-581.

Author's address:

Prof. Andre' I. Khuri
Department of Statistics
Nuclear Sciences Center
The University of Florida
Gainesville, Florida 32611
U.S.A.

DTIC
END
4-86